PURDUE UNIVERSITY

SC22 Dallas, TX | hpc accelerates.

Containerized Bioinformatics Ecosystem for HPC

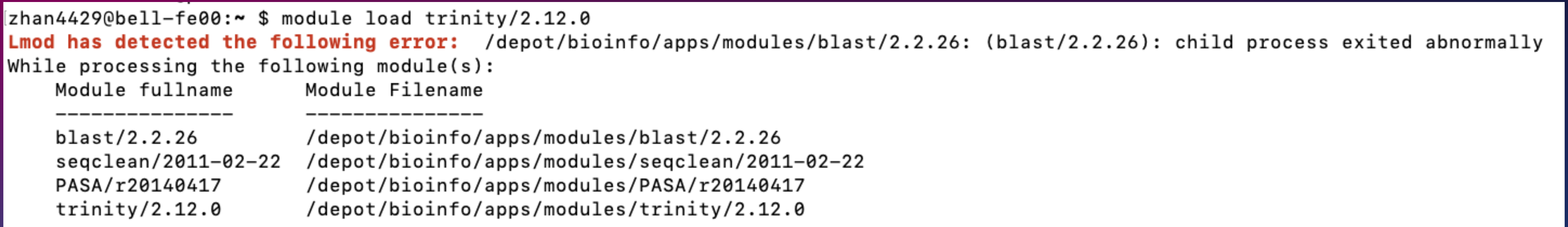
Yucheng Zhang[1], Lev Gorenstein[1], Payas Bhutra[2], Ryan DeRue[1]

1. Rosen Center for Advanced Computing, Purdue University, West Lafayette, IN, USA

2. Department of Computer Science, Purdue University, West Lafayette, IN, USA

Rationale

- Purdue RCAC have to manage multiple production systems, including 6 community clusters and ACCESS Anvil.
- Purdue has a large number of biological researchers studying various areas, such as agriculture, ecology, animal science, health science, etc.



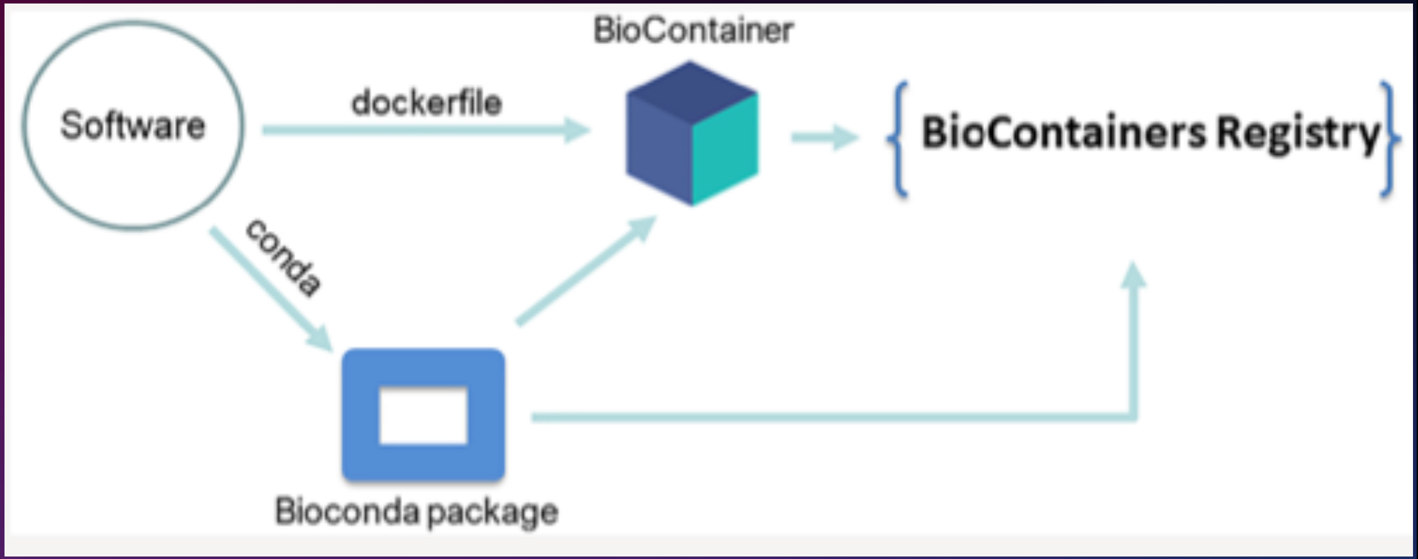
```
zhan4429@bell-fe00:~ $ module load trinity/2.12.0
Lmod has detected the following error:  /depot/bioinfo/apps/modules/blast/2.2.26: (blast/2.2.26): child process exited abnormally
While processing the following module(s):
    Module fullname       Module Filename
    ---------------       ---------------
    blast/2.2.26          /depot/bioinfo/apps/modules/blast/2.2.26
    seqclean/2011-02-22   /depot/bioinfo/apps/modules/seqclean/2011-02-22
    PASA/r20140417        /depot/bioinfo/apps/modules/PASA/r20140417
    trinity/2.12.0        /depot/bioinfo/apps/modules/trinity/2.12.0
```

An easy and reliable approach to manage a large stack of bioinformatics applications is urgently needed.

BioContainers

BIOCONDA®

- BioContainers is integrated with Bioconda, which is the conda channel for bioinformatics applications.
- BioContainers registry is the largest registry for bioinformatics applications.
- As of today, BioContainers provides containers for over 10 thousand bioinformatics applications.



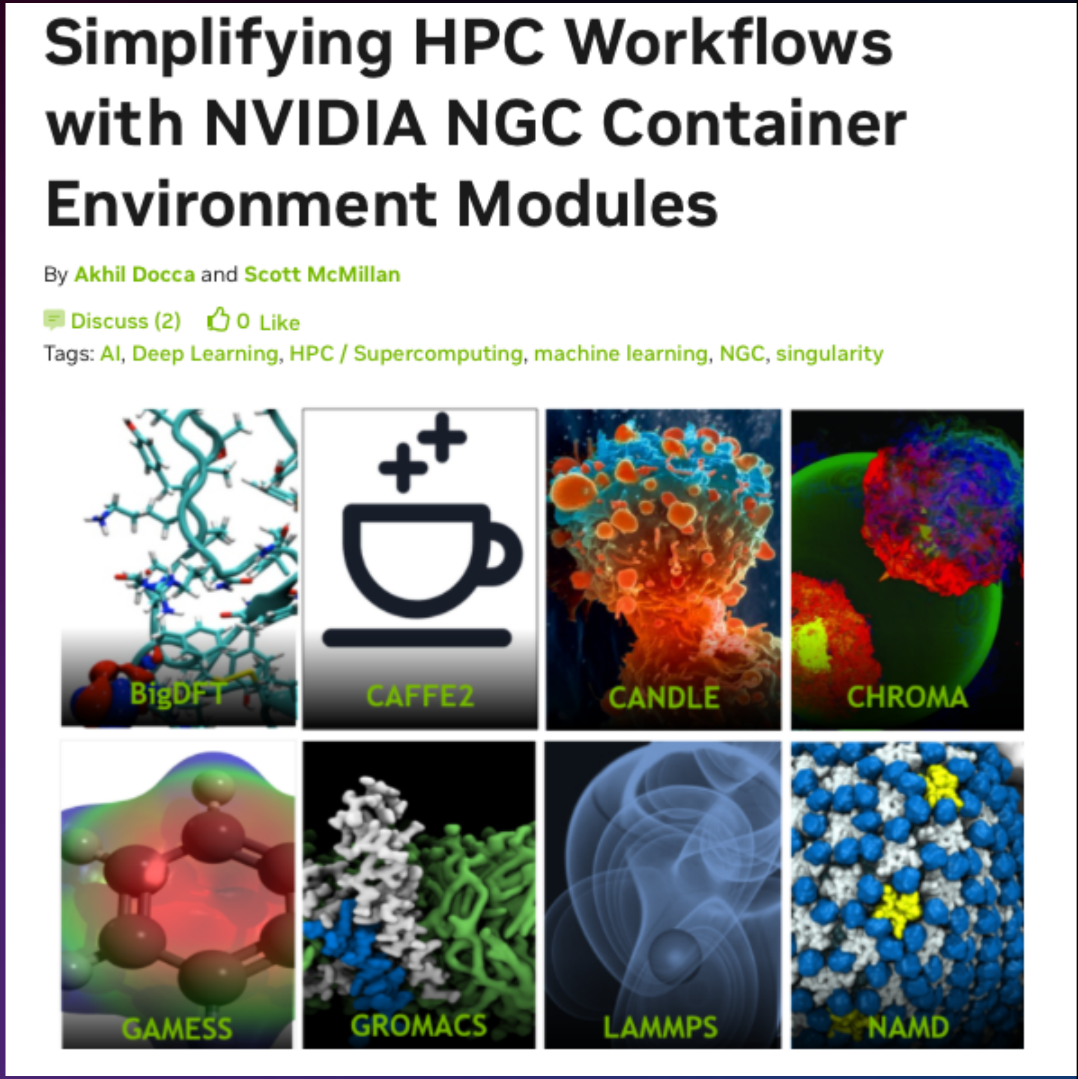
J. Proteome Res. 2021, 20, 4, 2056–2061

NGC container environment modules

NGC container environment modules are lightweight wrappers that make it possible to transparently use NGC containers as environment modules.

1. Allow HPC users to utilize familiar environment module commands.
2. Leverage all the benefits of containers, including portability and reproducibility.

https://github.com/NVIDIA/ngc-container-environment-modules



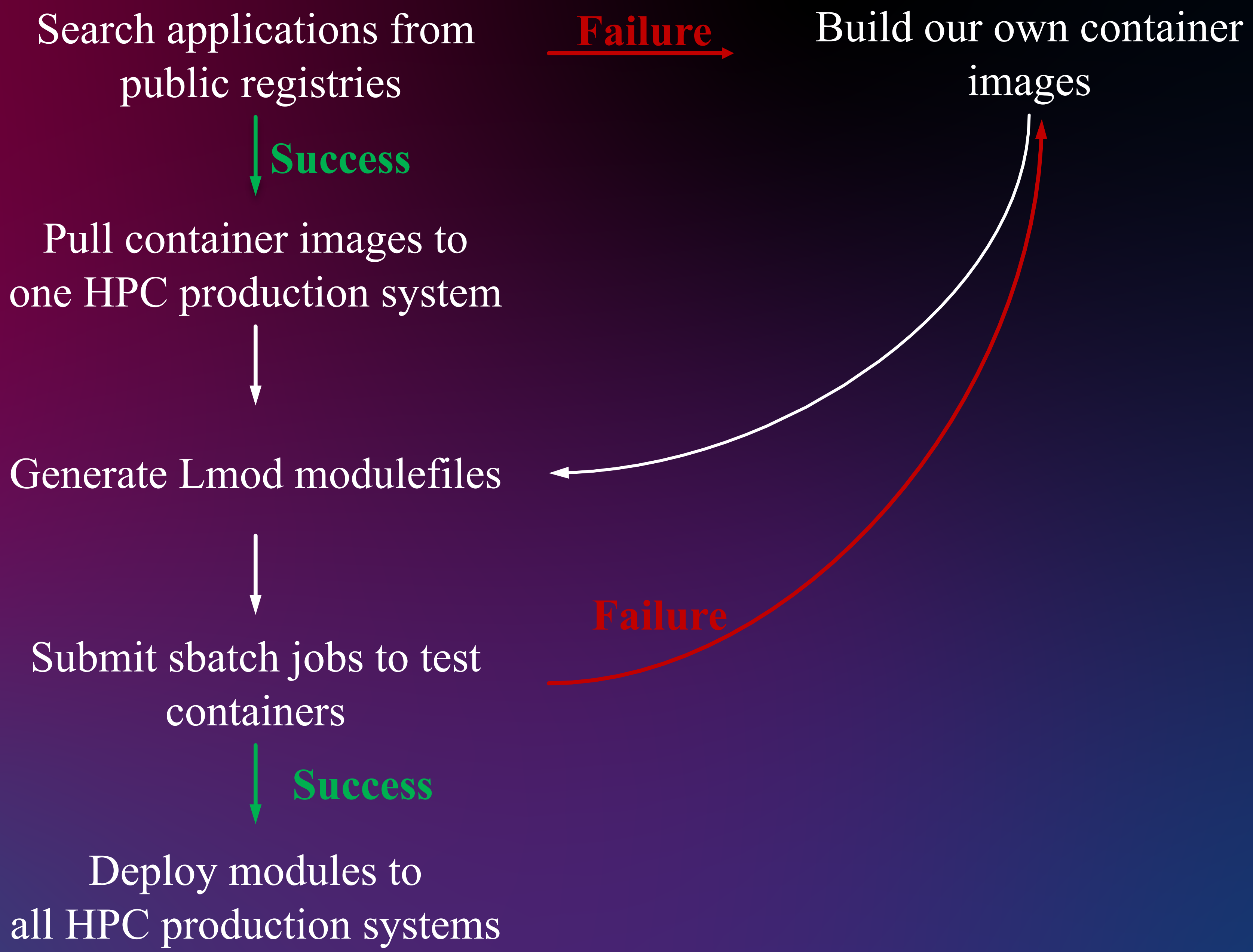
Simplifying HPC Workflows with NVIDIA NGC Container Environment Modules

By Akhil Docca and Scott McMillan

Discuss (2) 0 Like

Tags: AI, Deep Learning, HPC / Supercomputing, machine learning, NGC, singularity

Pull/build, test before deployment



Search applications from public registries

Failure

Build our own container images

Success

Pull container images to one HPC production system

Generate Lmod modulefiles

Submit sbatch jobs to test containers

Failure

Success

Deploy modules to all HPC production systems

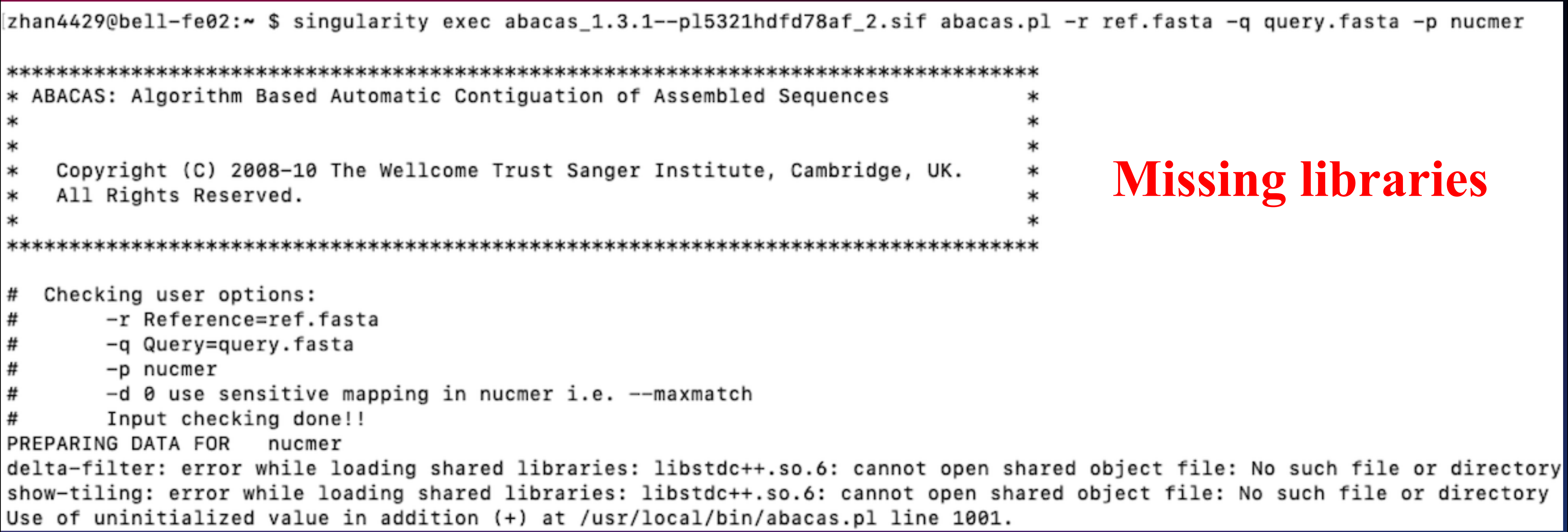
Pulling images and generating Lmod modulefiles

1. **bioc_pull2sif.sh**
   - a wrapper around “singularity pull”
   - Outputing image names following the convention set by NGC container environment modules
2. **bioc_pull2mod.sh**
   - Generate Lmod modulefile
3. **bioc_pull_and_module.sh**
   - A wrapper combining the first two scripts
   - Given a container URI, it will pull the container image and generate its modulefile

Special Lmod modulefile setup

1. **Add help/whatis information**
2. **GUI applications:** bind X11 session information in ThinLinc
    - append_path("SINGULARITY_BIND", "/var/opt", ",")
    - append_path("SINGULARITY_BIND", "/run/user", ",")
3. **Environment variables:** environment variables associated with location to database or config files
    - pushenv("NAME", "value") ## set variable in host
    - pushenv("SINGULARITYENV_NAME", "value") ## set variable inside container
4. **Adding executable path to PATH**
    - pushenv("SINGULARITYENV_PREPEND_PATH", "/path/to/pkg/bin")
5. **Bind paths**: bind database or config files
    - append_path("SINGULARITY_BIND", "hostdir:containerdir", ",")

Testing modules before deployment

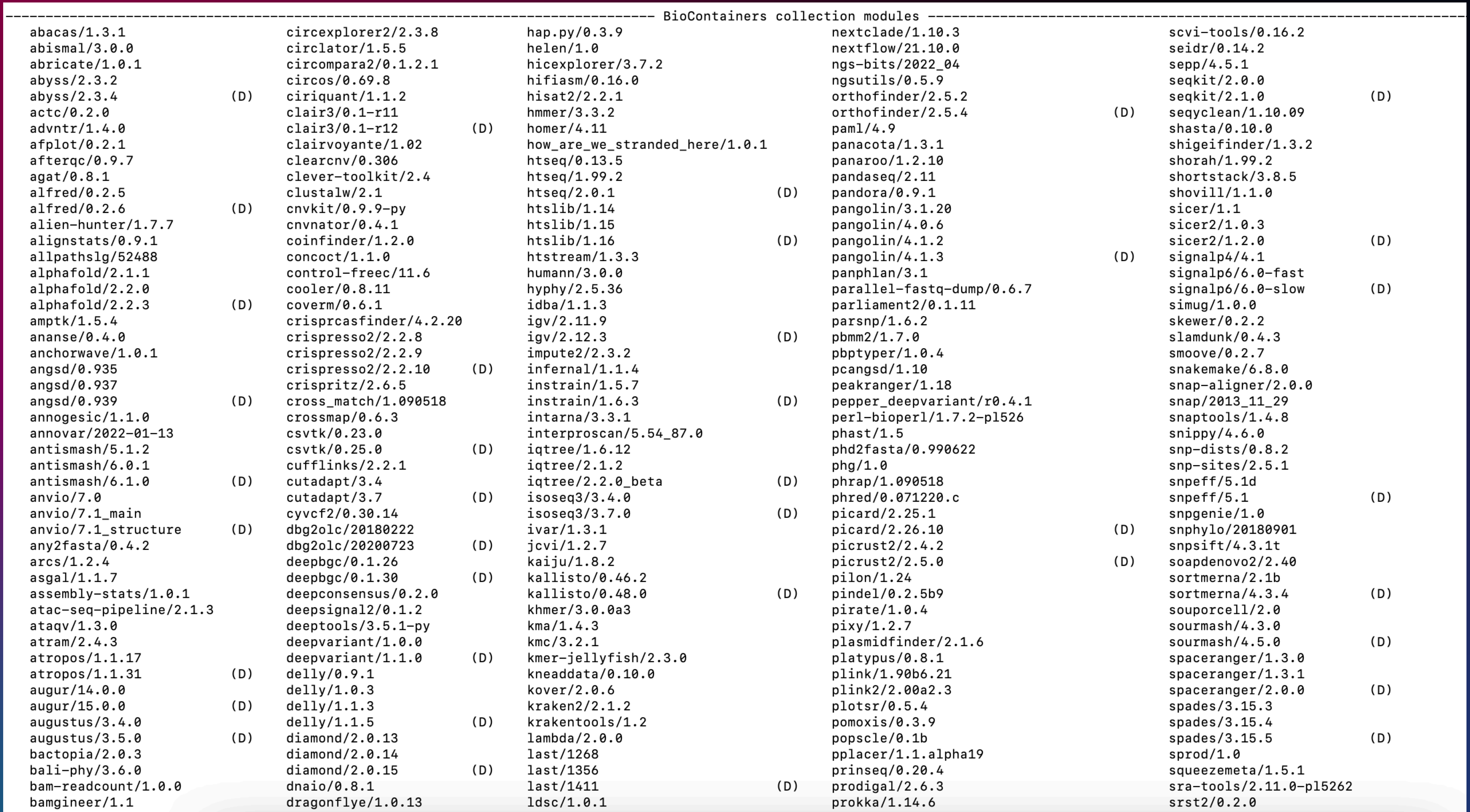


```
[zhan4429@bell-fe02:~ $ singularity exec abacas_1.3.1--pl5321hdfd78af_2.sif abacas.pl -r ref.fasta -q query.fasta -p nucmer

************************************************************************************
* ABACAS: Algorithm Based Automatic Contiguation of Assembled Sequences            *
*                                                                                  *
*                                                                                  *
*   Copyright (C) 2008-10 The Wellcome Trust Sanger Institute, Cambridge, UK.      *
*   All Rights Reserved.                                                           *
*                                                                                  *
************************************************************************************

#  Checking user options:
#       -r Reference=ref.fasta
#       -q Query=query.fasta
#       -p nucmer
#       -d 0 use sensitive mapping in nucmer i.e. --maxmatch
#       Input checking done!!
PREPARING DATA FOR   nucmer
delta-filter: error while loading shared libraries: libstdc++.so.6: cannot open shared object file: No such file or directory
show-tiling: error while loading shared libraries: libstdc++.so.6: cannot open shared object file: No such file or directory
Use of uninitialized value in addition (+) at /usr/local/bin/abacas.pl line 1001.
```

Missing libraries

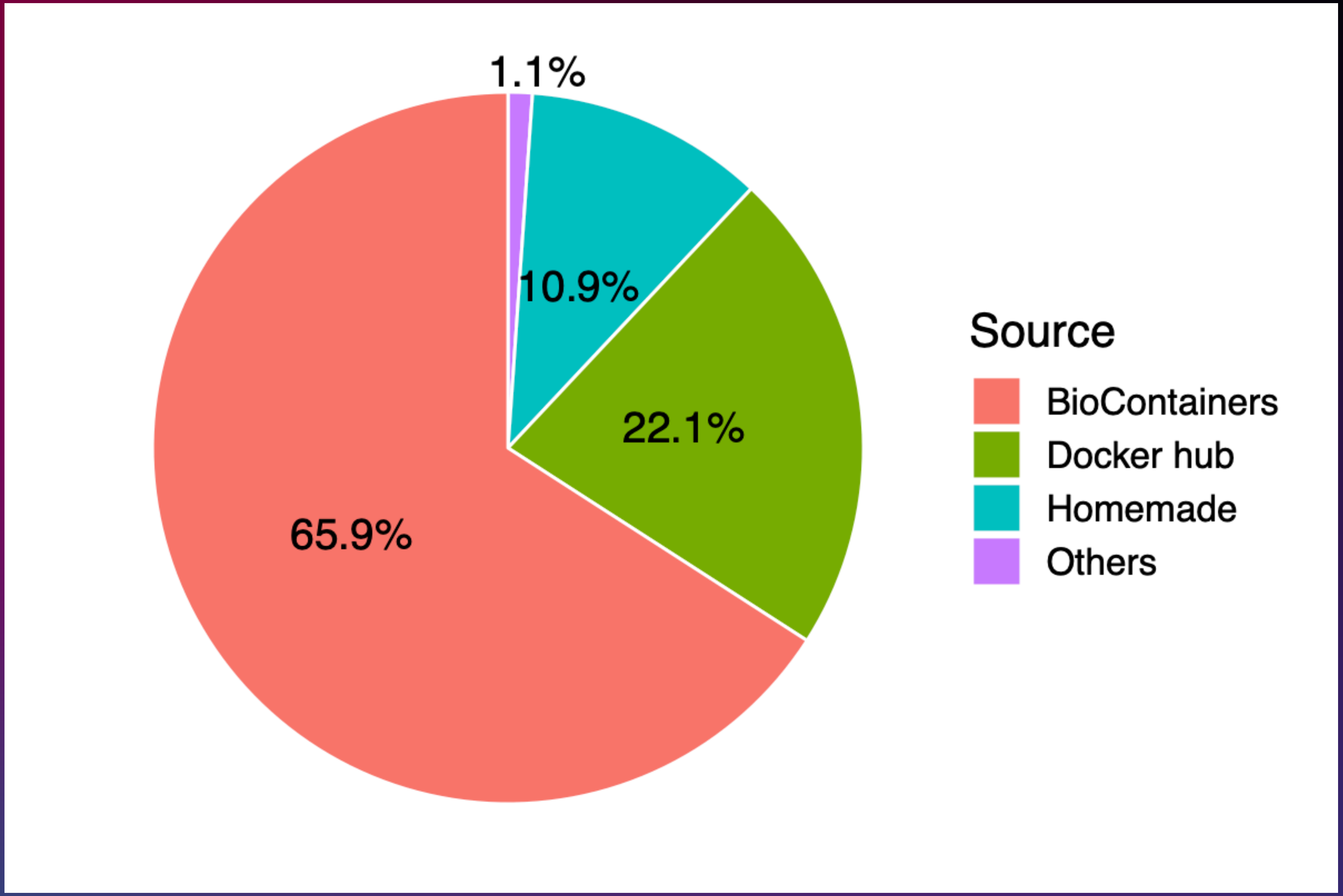
Running analysis with real-world datasets is the only reliable way to make sure applications work as expected.

600 modules for 500 applications



```
--------------------------------------------------------------------------------- BioContainers collection modules -------------------------------------------------------------------
   abacas/1.3.1                    circexplorer2/2.3.8           hap.py/0.3.9                          nextclade/1.10.3                          scvi-tools/0.16.2
   abismal/3.0.0                   circlator/1.5.5               helen/1.0                             nextflow/21.10.0                          seidr/0.14.2
   abricate/1.0.1                  circompara2/0.1.2.1           hicexplorer/3.7.2                     ngs-bits/2022_04                          sepp/4.5.1
   abyss/2.3.2                     circos/0.69.8                 hifiasm/0.16.0                        ngsutils/0.5.9                            seqkit/2.0.0
   abyss/2.3.4              (D)    ciriquant/1.1.2               hisat2/2.2.1                          orthofinder/2.5.2                         seqkit/2.1.0              (D)
   actc/0.2.0                      clair3/0.1-r11                hmmer/3.3.2                           orthofinder/2.5.4                  (D)    seqyclean/1.10.09
   advntr/1.4.0                    clair3/0.1-r12         (D)    homer/4.11                            paml/4.9                                  shasta/0.10.0
   afplot/0.2.1                    clairvoyante/1.02             how_are_we_stranded_here/1.0.1        panacota/1.3.1                            shigeifinder/1.3.2
   afterqc/0.9.7                   clearcnv/0.306                htseq/0.13.5                          panaroo/1.2.10                            shorah/1.99.2
   agat/0.8.1                      clever-toolkit/2.4            htseq/1.99.2                          pandaseq/2.11                             shortstack/3.8.5
   alfred/0.2.5                    clustalw/2.1                  htseq/2.0.1                    (D)    pandora/0.9.1                             shovill/1.1.0
   alfred/0.2.6             (D)    cnvkit/0.9.9-py               htslib/1.14                           pangolin/3.1.20                           sicer/1.1
   alien-hunter/1.7.7              cnvnator/0.4.1                htslib/1.15                           pangolin/4.0.6                            sicer2/1.0.3
   alignstats/0.9.1                coinfinder/1.2.0              htslib/1.16                    (D)    pangolin/4.1.2                            sicer2/1.2.0              (D)
   allpathslg/52488                concoct/1.1.0                 htstream/1.3.3                        pangolin/4.1.3                     (D)    signalp4/4.1
   alphafold/2.1.1                 control-freec/11.6            humann/3.0.0                          panphlan/3.1                              signalp6/6.0-fast
   alphafold/2.2.0                 cooler/0.8.11                 hyphy/2.5.36                          parallel-fastq-dump/0.6.7                 signalp6/6.0-slow         (D)
   alphafold/2.2.3          (D)    coverm/0.6.1                  idba/1.1.3                            parliament2/0.1.11                        simug/1.0.0
   amptk/1.5.4                     crisprcasfinder/4.2.20        igv/2.11.9                            parsnp/1.6.2                              skewer/0.2.2
   ananse/0.4.0                    crispresso2/2.2.8             igv/2.12.3                     (D)    pbmm2/1.7.0                               slamdunk/0.4.3
   anchorwave/1.0.1                crispresso2/2.2.9             impute2/2.3.2                         pbptyper/1.0.4                            smoove/0.2.7
   angsd/0.935                     crispresso2/2.2.10     (D)    infernal/1.1.4                        pcangsd/1.10                              snakemake/6.8.0
   angsd/0.937                     crispritz/2.6.5               instrain/1.5.7                        peakranger/1.18                           snap-aligner/2.0.0
   angsd/0.939              (D)    cross_match/1.090518          instrain/1.6.3                 (D)    pepper_deepvariant/r0.4.1                 snap/2013_11_29
   annogesic/1.1.0                 crossmap/0.6.3                intarna/3.3.1                         perl-bioperl/1.7.2-pl526                  snaptools/1.4.8
   annovar/2022-01-13              csvtk/0.23.0                  interproscan/5.54_87.0                phast/1.5                                 snippy/4.6.0
   antismash/5.1.2                 csvtk/0.25.0           (D)    iqtree/1.6.12                         phd2fasta/0.990622                        snp-dists/0.8.2
   antismash/6.0.1                 cufflinks/2.2.1               iqtree/2.1.2                          phg/1.0                                   snp-sites/2.5.1
   antismash/6.1.0          (D)    cutadapt/3.4                  iqtree/2.2.0_beta              (D)    phrap/1.090518                            snpeff/5.1d
   anvio/7.0                       cutadapt/3.7           (D)    isoseq3/3.4.0                         phred/0.071220.c                          snpeff/5.1                (D)
   anvio/7.1_main                  cyvcf2/0.30.14                isoseq3/3.7.0                  (D)    picard/2.25.1                             snpgenie/1.0
   anvio/7.1_structure      (D)    dbg2olc/20180222              ivar/1.3.1                            picard/2.26.10                     (D)    snphylo/20180901
   any2fasta/0.4.2                 dbg2olc/20200723       (D)    jcvi/1.2.7                            picrust2/2.4.2                            snpsift/4.3.1t
   arcs/1.2.4                      deepbgc/0.1.26                kaiju/1.8.2                           picrust2/2.5.0                     (D)    soapdenovo2/2.40
   asgal/1.1.7                     deepbgc/0.1.30         (D)    kallisto/0.46.2                       pilon/1.24                                sortmerna/2.1b
   assembly-stats/1.0.1            deepconsensus/0.2.0           kallisto/0.48.0                (D)    pindel/0.2.5b9                            sortmerna/4.3.4           (D)
   atac-seq-pipeline/2.1.3         deepsignal2/0.1.2             khmer/3.0.0a3                         pirate/1.0.4                              souporcell/2.0
   ataqv/1.3.0                     deeptools/3.5.1-py            kma/1.4.3                             pixy/1.2.7                                sourmash/4.3.0
   atram/2.4.3                     deepvariant/1.0.0             kmc/3.2.1                             plasmidfinder/2.1.6                       sourmash/4.5.0            (D)
   atropos/1.1.17                  deepvariant/1.1.0      (D)    kmer-jellyfish/2.3.0                  platypus/0.8.1                            spaceranger/1.3.0
   atropos/1.1.31           (D)    delly/0.9.1                   kneaddata/0.10.0                      plink/1.90b6.21                           spaceranger/1.3.1
   augur/14.0.0                    delly/1.0.3                   kover/2.0.6                           plink2/2.00a2.3                           spaceranger/2.0.0         (D)
   augur/15.0.0             (D)    delly/1.1.3                   kraken2/2.1.2                         plotsr/0.5.4                              spades/3.15.3
   augustus/3.4.0                  delly/1.1.5            (D)    krakentools/1.2                       pomoxis/0.3.9                             spades/3.15.4
   augustus/3.5.0           (D)    diamond/2.0.13                lambda/2.0.0                          popscle/0.1b                              spades/3.15.5             (D)
   bactopia/2.0.3                  diamond/2.0.14                last/1268                             pplacer/1.1.alpha19                       sprod/1.0
   bali-phy/3.6.0                  diamond/2.0.15         (D)    last/1356                             prinseq/0.20.4                            squeezemeta/1.5.1
   bam-readcount/1.0.0             dnaio/0.8.1                   last/1411                      (D)    prodigal/2.6.3                            sra-tools/2.11.0-pl5262
   bamgineer/1.1                   dragonflye/1.0.13             ldsc/1.0.1                            prokka/1.14.6                             srst2/0.2.0
```

Sources of container images

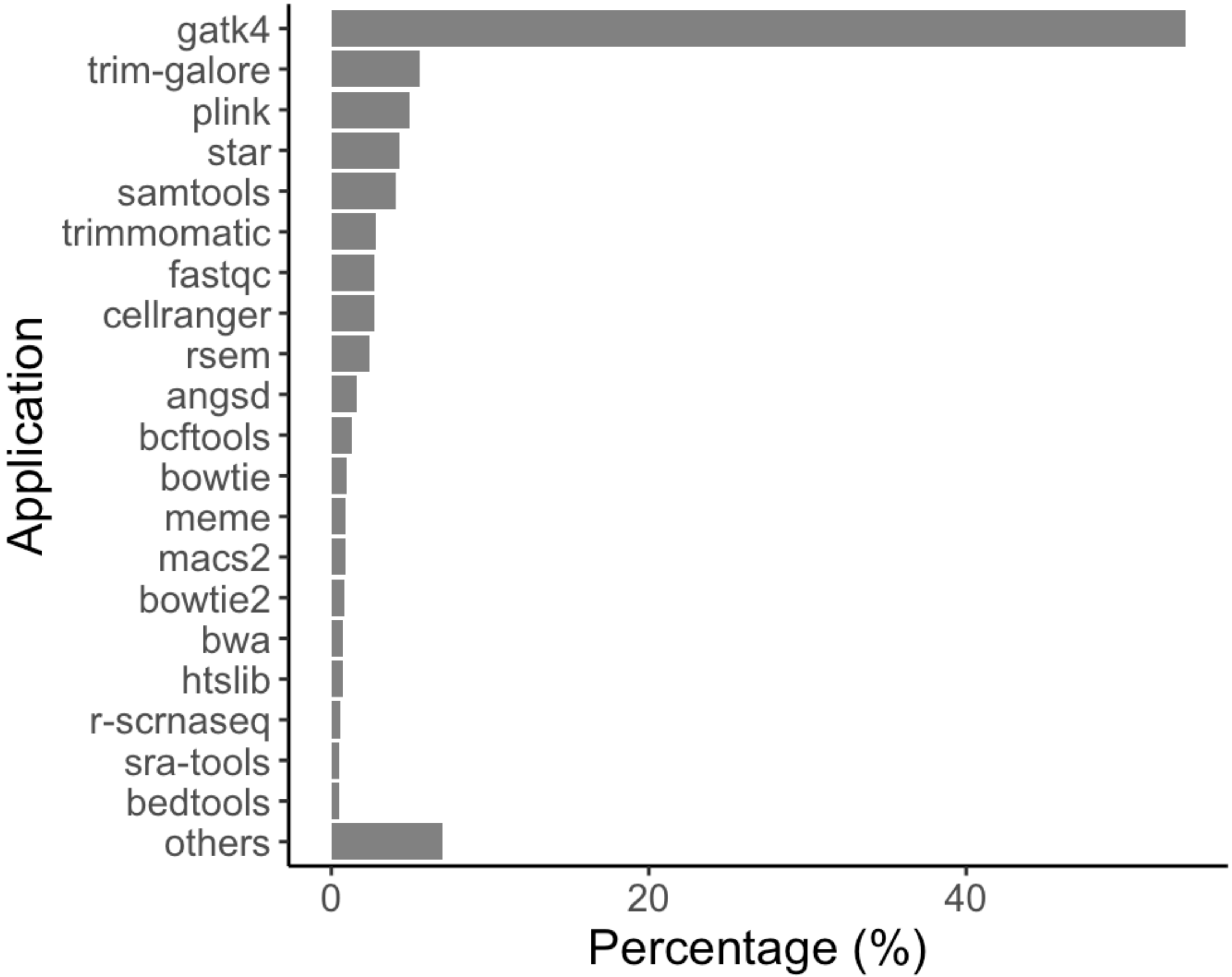


Bioinformatics applications on HPC

1. **Fewer than 100**: most HPC centers
2. **~1000**: clusters designed for biosciences, e.g. NIH’s Biowulf and Cornell’s BioHPC
3. **Thousands**: the whole BioContainers project in TACC’s clusters

   RollingGantryCrane (https://github.com/TACC/rgc)

Too many is not necessary



Adding Jupyter support

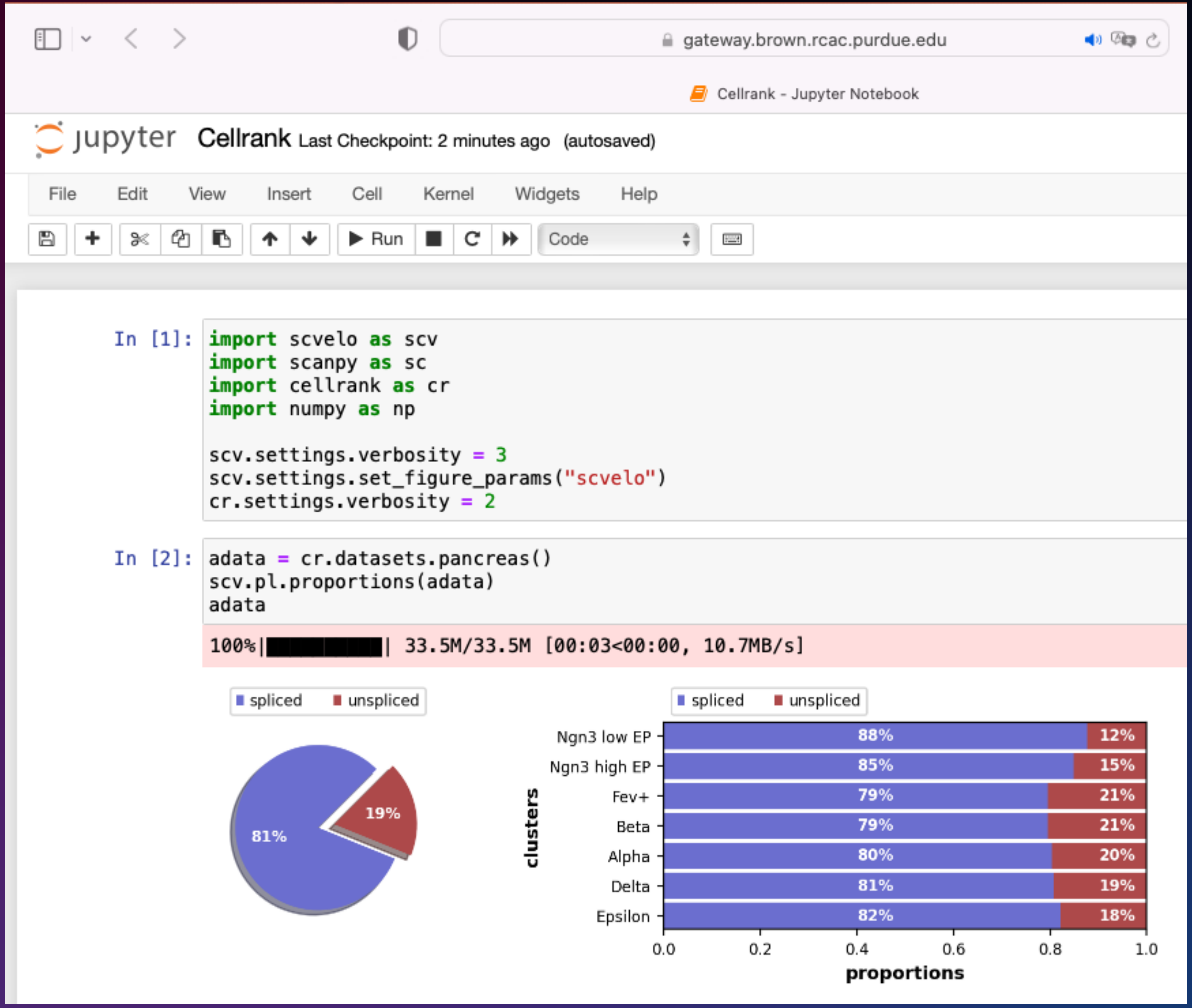
```
Bootstrap: docker
From: quay.io/biocontainers/cellrank:1.5.1--pyhdfd78af_0

%labels
    Author:  Yucheng Zhang zhan4429@purdue.edu
    Version:    1.5.1

%help
   CellRank with Jupyter support.

%post
   pip install ipython
   pip install ipykernel
```

Hopefully BioContainers developers can install ipython and ipykernel into all python-based container images.

Open OnDemand Jupyter

```
{
 "argv": [
  "/usr/bin/singularity",
  "exec",
  "/apps/biocontainers/images/cellrank_1.5.1.sif
",
  "python",
  "-m",
  "ipykernel_launcher",
  "-f",
  "{connection_file}"
 ],
 "display_name": "Cellrank",
 "language": "python"
}
```

$HOME/.local/share/jupyter/kernels/



```
In [1]: import scvelo as scv
        import scanpy as sc
        import cellrank as cr
        import numpy as np

        scv.settings.verbosity = 3
        scv.settings.set_figure_params("scvelo")
        cr.settings.verbosity = 2
```

```
In [2]: adata = cr.datasets.pancreas()
        scv.pl.proportions(adata)
        adata
```

```
100%|██████████| 33.5M/33.5M [00:03<00:00, 10.7MB/s]
```

Bundle applications into a single container image

With containers, it is easy to install not just a single application, but also bundles and collections of multiple applications working in concert and dedicated to a specific research workflow.

R-RNAseq

Customized R container for RNAseq analysis.

- ComplexHeatmap
- DESeq2
- DEXSeq
- edgeR
- ggrepel
- Limma
- pheatmap
- tidyverse

R-scRNAseq

Customized R container for scRNAseq analysis.

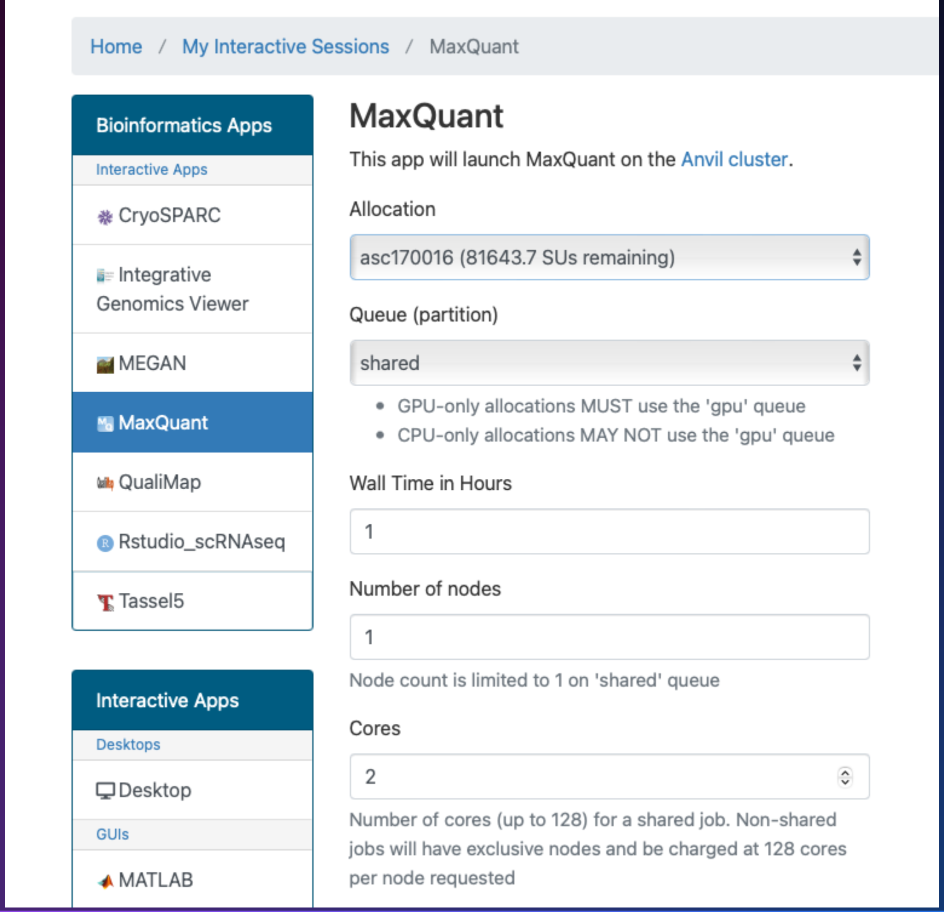
- CoGAPS
- DESeq2
- doSNOW
- DropletUtils
- edgeR
- Limma
- miQC
- monocle
- monocle3
- Nebulosa
- rliger
- scCATCH
- scDblFinder
- SCHNAPPs
- scMappR
- seurat
- seurat-wrappers
- SingleR
- SnapATAC
- SoupX
- tidyverse
- tricycle
- velocyto.R

And more

Open OnDemand

- For bioinformatics applications that use a native graphics user interface (GUI) and that have a large computational or memory footprint, we employ Open OnDemand to allow users to easily allocate appropriate amount of resources and submit jobs through a convenient web interface.

- We create a simple workflow for rapid deployment of containers to Open OnDemand of any cluster.
  1. default_biocontainer_template: a template Open OnDemand application directory for a generic VNC desktop application.
  2. deploy_biocontainer: a helper script that makes a copy of the template directory and performs the necessary substitutions to the relevant files.



MaxQuant

This app will launch MaxQuant on the Anvil cluster.

Allocation: asc170016 (81643.7 SUs remaining)

Queue (partition): shared

- GPU-only allocations MUST use the 'gpu' queue
- CPU-only allocations MAY NOT use the 'gpu' queue

Wall Time in Hours: 1

Number of nodes: 1

Node count is limited to 1 on 'shared' queue

Cores: 2

Number of cores (up to 128) for a shared job. Non-shared jobs will have exclusive nodes and be charged at 128 cores per node requested

Turning containers into OOD aps

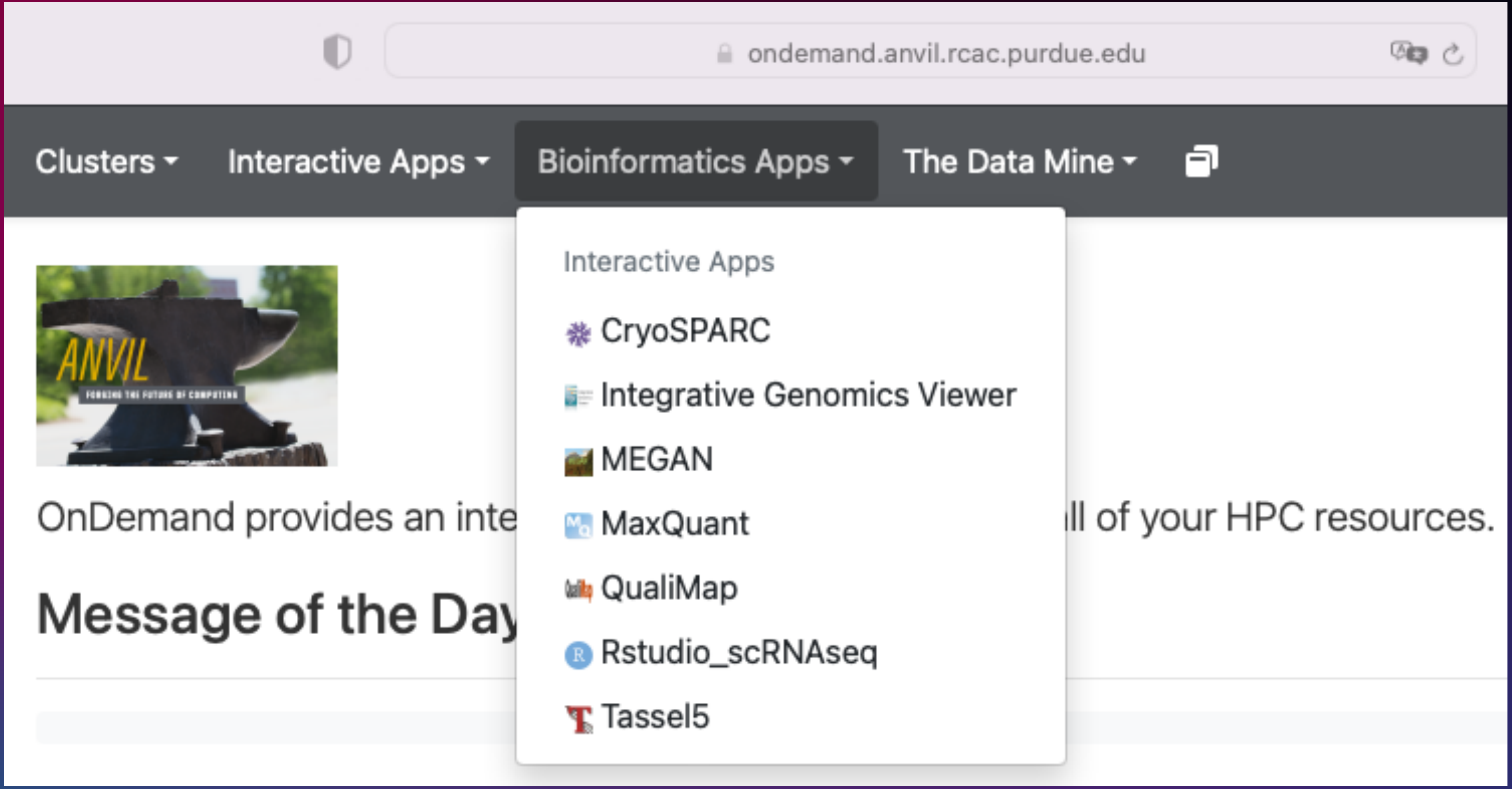
With the template directory default_biocontainer_template and the deploy_biocontainer script, we can easily turn container images into Open OnDemand interactive applications using a one-line command:

```
deploy_biocontainer  --name appName \
                     --directory folderName \
                     --image app.sif  \
                     --command launchCommand \
                     default_biocontainer_template
```

```
ENTRY="[Desktop Entry]
Type=Application Name=IGV
Comment=
Exec=/bin/bash -lc \"singularity exec /apps/biocontainers/images/igv_2.12.3.sif igv.sh\"
Path=
Terminal=false
StartupNotify=false
Categories=Cluster"
echo -e "$ENTRY" > "${AUTOSTART}/igv.desktop"
```

A snippet from xfce.sh for the genomic browser IGV

Open OnDemand Applications



ondemand.anvil.rcac.purdue.edu

OnDemand provides an inte ll of your HPC resources.

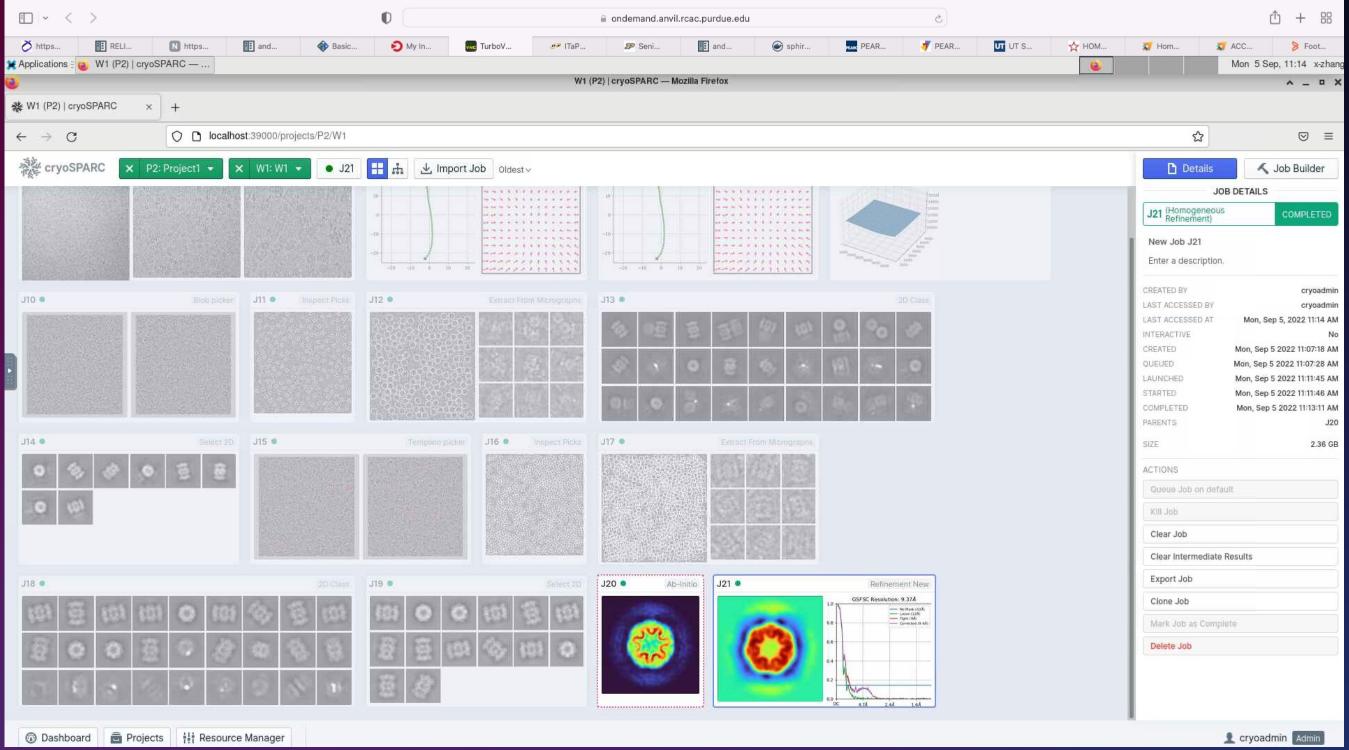
Message of the Day

Not only bioinformatics

cryoSPARC

RELION

crYOLO



JOB DETAILS

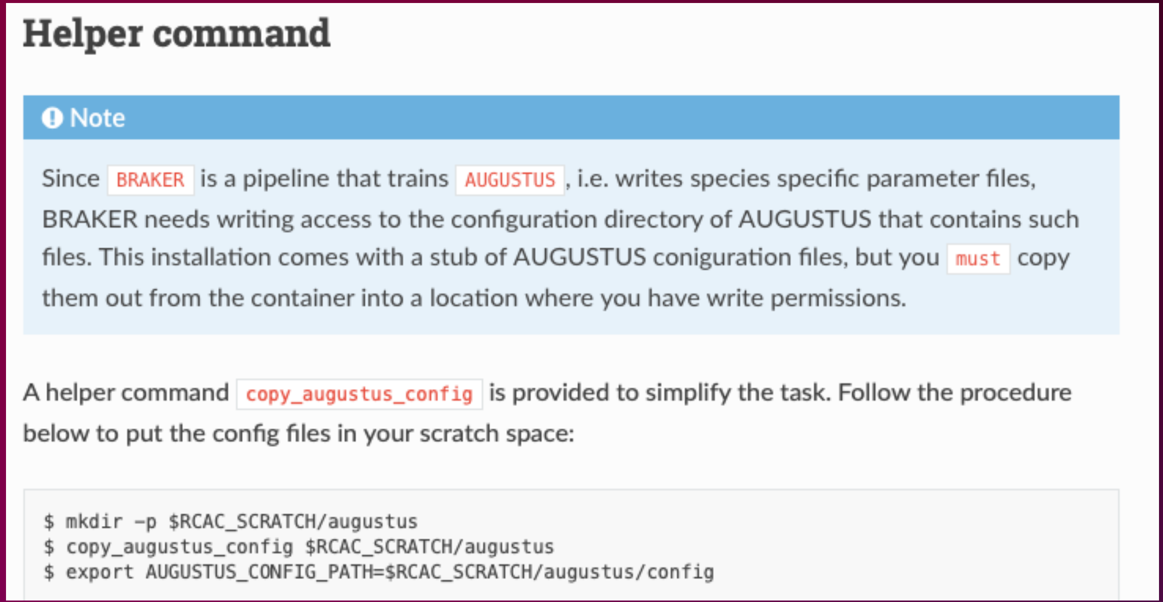
J21 (Homogeneous Refinement) COMPLETED

New Job J21

Enter a description.

CREATED BY	cryoadmin
LAST ACCESSED BY	cryoadmin
LAST ACCESSED AT	Mon, Sep 5, 2022 11:14 AM
INTERACTIVE	No
CREATED	Mon, Sep 5 2022 11:07:18 AM
QUEUED	Mon, Sep 5 2022 11:07:28 AM
LAUNCHED	Mon, Sep 5 2022 11:11:45 AM
STARTED	Mon, Sep 5 2022 11:11:46 AM
COMPLETED	Mon, Sep 5 2022 11:13:11 AM
PARENTS	J20
SIZE	2.36 GB

Thank Michael Dickens from Texas A&M University for showing us how to build and run the cryoSPARC container.



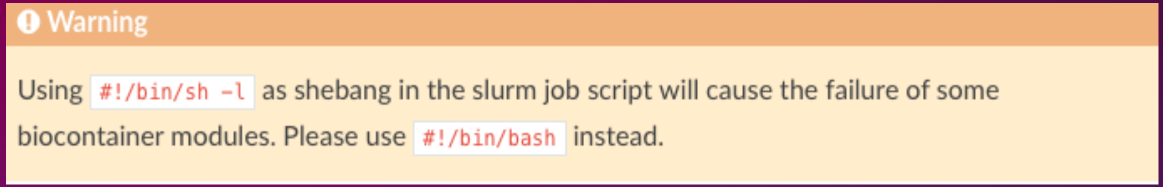
Helper command

Note

Since `BRAKER` is a pipeline that trains `AUGUSTUS`, i.e. writes species specific parameter files, BRAKER needs writing access to the configuration directory of AUGUSTUS that contains such files. This installation comes with a stub of AUGUSTUS coniguration files, but you `must` copy them out from the container into a location where you have write permissions.

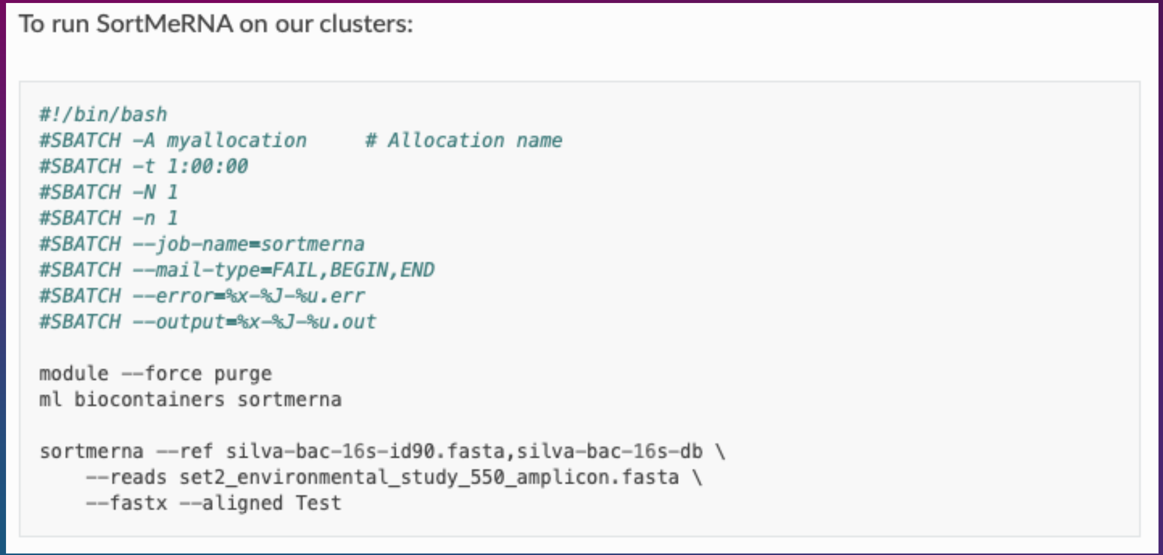
A helper command `copy_augustus_config` is provided to simplify the task. Follow the procedure below to put the config files in your scratch space:

```
$ mkdir -p $RCAC_SCRATCH/augustus
$ copy_augustus_config $RCAC_SCRATCH/augustus
$ export AUGUSTUS_CONFIG_PATH=$RCAC_SCRATCH/augustus/config
```



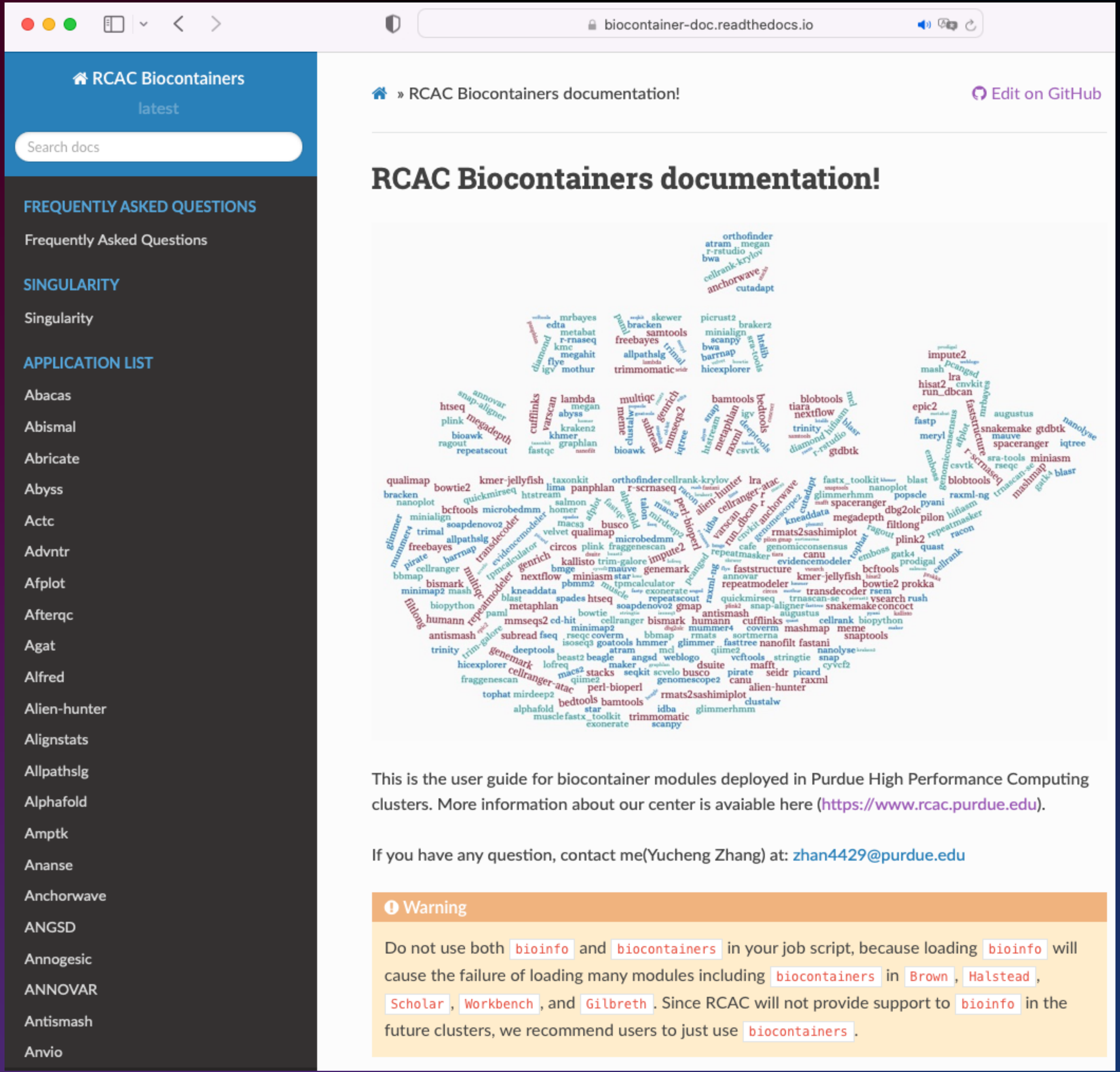
Warning

Using `#!/bin/sh -l` as shebang in the slurm job script will cause the failure of some biocontainer modules. Please use `#!/bin/bash` instead.



To run SortMeRNA on our clusters:

```
#!/bin/bash
#SBATCH -A myallocation     # Allocation name
#SBATCH -t 1:00:00
#SBATCH -N 1
#SBATCH -n 1
#SBATCH --job-name=sortmerna
#SBATCH --mail-type=FAIL,BEGIN,END
#SBATCH --error=%x-%J-%u.err
#SBATCH --output=%x-%J-%u.out

module --force purge
ml biocontainers sortmerna

sortmerna --ref silva-bac-16s-id90.fasta,silva-bac-16s-db \
    --reads set2_environmental_study_550_amplicon.fasta \
    --fastx --aligned Test
```



RCAC Biocontainers documentation!

This is the user guide for biocontainer modules deployed in Purdue High Performance Computing clusters. More information about our center is avaiable here (https://www.rcac.purdue.edu).

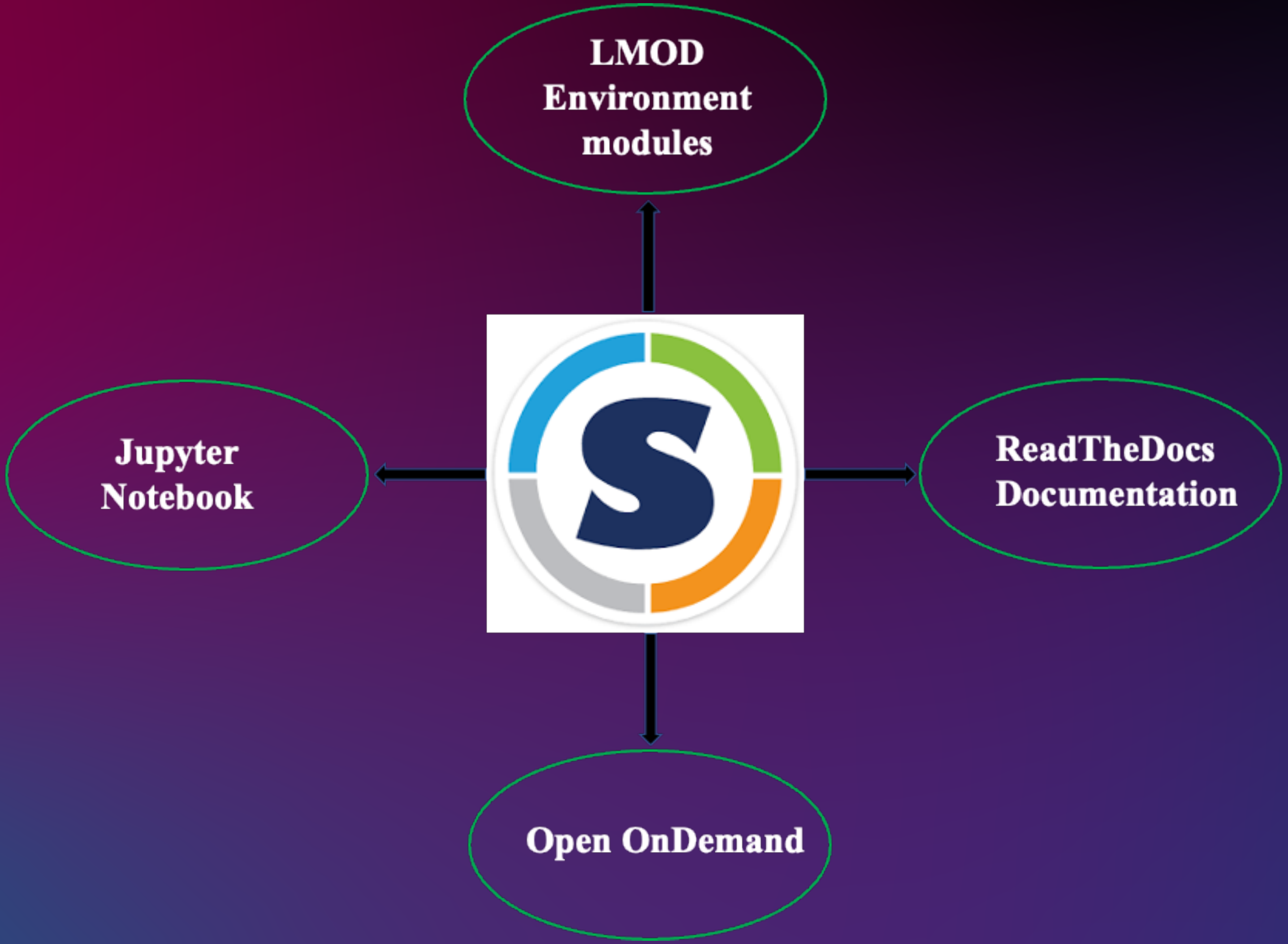
If you have any question, contact me(Yucheng Zhang) at: zhan4429@purdue.edu

Warning

Do not use both `bioinfo` and `biocontainers` in your job script, because loading `bioinfo` will cause the failure of loading many modules including `biocontainers` in `Brown`, `Halstead`, `Scholar`, `Workbench`, and `Gilbreth`. Since RCAC will not provide support to `bioinfo` in the future clusters, we recommend users to just use `biocontainers`.

https://biocontainer-doc.readthedocs.io/en/latest/

Containerized Bioinformatics Ecosystem



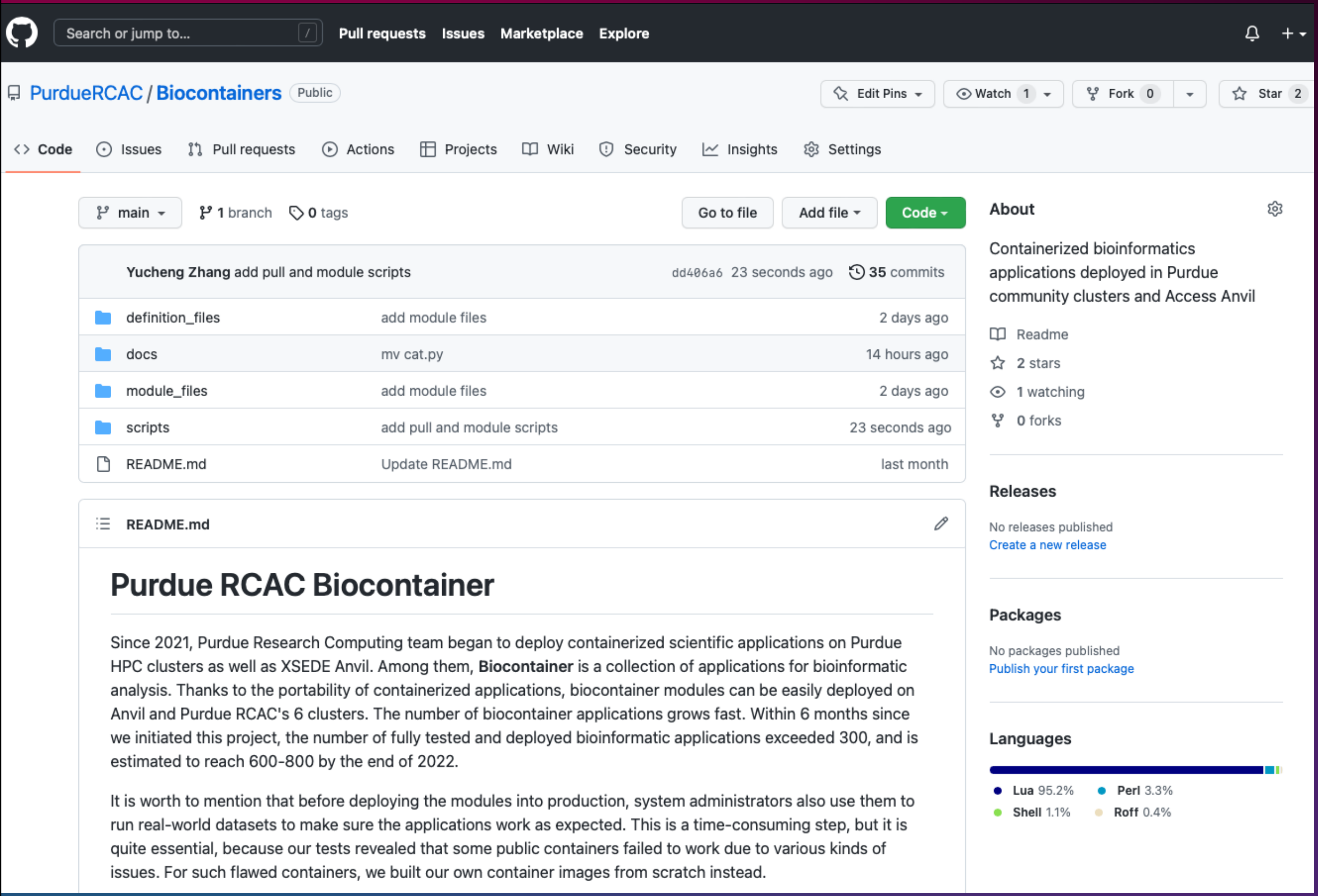
LMOD Environment modules

Jupyter Notebook

ReadTheDocs Documentation

Open OnDemand

Interested in building a similar ecosystem in your center?



Yucheng Zhang add pull and module scripts dd406a6 23 seconds ago 35 commits

definition_files	add module files	2 days ago
docs	mv cat.py	14 hours ago
module_files	add module files	2 days ago
scripts	add pull and module scripts	23 seconds ago
README.md	Update README.md	last month

README.md

Purdue RCAC Biocontainer

Since 2021, Purdue Research Computing team began to deploy containerized scientific applications on Purdue HPC clusters as well as XSEDE Anvil. Among them, **Biocontainer** is a collection of applications for bioinformatic analysis. Thanks to the portability of containerized applications, biocontainer modules can be easily deployed on Anvil and Purdue RCAC's 6 clusters. The number of biocontainer applications grows fast. Within 6 months since we initiated this project, the number of fully tested and deployed bioinformatic applications exceeded 300, and is estimated to reach 600-800 by the end of 2022.

It is worth to mention that before deploying the modules into production, system administrators also use them to run real-world datasets to make sure the applications work as expected. This is a time-consuming step, but it is quite essential, because our tests revealed that some public containers failed to work due to various kinds of issues. For such flawed containers, we built our own container images from scratch instead.

About

Containerized bioinformatics applications deployed in Purdue community clusters and Access Anvil

Readme
2 stars
1 watching
0 forks

Releases

No releases published
Create a new release

Packages

No packages published
Publish your first package

Languages

● Lua 95.2%	● Perl 3.3%
● Shell 1.1%	● Roff 0.4%

Contributions are welcome!

- If you find issues or bugs, please open an issue in the GitHub repository.
- Our goal is to improve the Biocontainer project together with all centers, and we need your ideas and input to keep on improving.
- We welcome any contribution including scripts, modulefiles, definition files, etc.

git clone https://github.com/PurdueRCAC/Biocontainers.git

Thank you!

Contributors

- Yucheng Zhang
- Lev Gorenstein
- Payas Bhutra
- Ryan DeRue

Purdue RCAC

- Preston Smith
- Xiao Zhu (Intel)
- Arman Pazouki
- Amiya Maji
- Geoffrey Lentner
- Guangzhen Jin
- Sarah Rodenbeck
- Steve Kelley
- Yang Hong
- Tsai-wei Wu
- Nannan Shan
- Eric Adams
- Ruyi Li

HUST22 Committee